**Revista Colombiana de Estadística**
January 2017, Volume 40, Issue 1, pp. 29 to 44
DOI: http://dx.doi.org/10.15446/rce.v40n1.53459

# Estimation of Sensitive Attributes Using a Stratified Kuk Randomization Device

## Estimación de atributos sensibles usando un mecanismo de aleatorización estratificado de Kuk

Lee Gi-Sung[1,a], Hong Ki-Hak[2,b], Kim Jong-Min[3,c], Son Chang-Kyoon[4,d]

[1]Department of Child Welfare, Woosuk University, Wanju-gun Jeonbuk, Korea

[2]Department of Computer Science, Dongshin University, Naju Jeonnam country,

[3]Statistics Discipline, Division of Science and Mathematics, University of Minnesota at Morris, Minnesota, USA

[4]Department of Applied Statistics, Dongguk University, Gyeongju Gyeongbuk, Republic of Korea

### Abstract

This paper suggests a stratified Kuk model to estimate the proportion of sensitive attributes of a population composed by a number of strata; this is undertaken by applying stratified sampling to the adjusted Kuk model. The paper estimates sensitive parameters when the size of the stratum is known by taking proportional and optimal allocation methods into account and then extends to the case of an unknown stratum size, estimating sensitive parameters by applying stratified double sampling and checking the two allocation methods. Finally, the paper compares the efficiency of the proposed model to that of the Su, Sedory and Singh model and the adjusted Kuk model in terms of the estimator variance.

***Key words*:** Adjusted Kuk Model, Randomized Response Model, Sensitive Attribute, Stratified Double Sampling, Stratified Sampling.

### Resumen

Este trabajo propone un modelo Kuk estratificado para estimar la proporción de atributos sensibles de una población compuesta por varios estratos mediante la aplicación de un muestreo estratificado al modelo Kuk ajustado.

[a]PhD. E-mail: gisung@woosuk.ac.kr
[b]PhD. E-mail: khhong@dsu.ac.kr
[c]PhD. E-mail: jongmink@mrs.umn.edu
[d]PhD. E-mail: ckson85@dongguk.ac.kr

El trabajo estima parámetros sensibles en el caso en que el tamaño del estrato es conocido mediante la adopción de métodos de asignación proporcionales y óptimos, y se extiende al caso de un tamaño de estrato desconocido, estimando parámetros sensibles mediante la aplicación de un doble muestreo estratificado y la comprobación de los dos métodos de asignación. Por último, el trabajo compara la eficiencia del modelo propuesto a la del modelo de Su, Sedory y Singh y el modelo Kuk ajustado en términos de la varianza del estimador.

***Palabras clave*:** modelo Kuk ajustado, modelo de respuesta aleatorizada, atributos sensibles, muestreo doble estratificado, muestreo estratificado.

# 1. Introduction

Warner (1965) was the first person to suggest an ingenious survey model called the randomized response model (RRM) to obtain sensitive information from respondents without disturbing their privacy by using a randomization device that contained the following two questions (a sensitive question and a nonsensitive one):

`Q1:`Do you have a sensitive attribute $A$? (with probability $P$),
`Q2:`Do you have a nonsensitive attribute $A^c$? (with probability $(1-P)$.

The probability of a "yes" answer is given by

$$\theta_W^* = P\pi + (1-P)(1-\pi). \tag{1}$$

Let $n\hat{\theta}_W^*$ be the number of "yes" responses in a sample of size $n$ respondents, and then the estimator $\hat{\pi}_W$ and the variance $V(\hat{\pi}_W)$ of its sensitive proportion $\pi$ are respectively

$$\hat{\pi}_W = \frac{\hat{\theta}_W^* - (1-P)}{2P-1}, P \neq 1/2, \tag{2}$$

$$V(\hat{\pi}_W) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2}. \tag{3}$$

Kuk (1990) suggested an RR model that makes use of two randomization devices. The first randomization device $R_1$ is composed of a deck of cards, and each card bears one of two possible questions with two possible outcomes:

`Q1:`Do you have a sensitive attribute $A$? (with probability $\theta_1$).
`Q2:`Do you have a nonsensitive attribute $A^c$? (with probability $1-\theta_1$).

The second randomization device $R_2$ is composed of a deck of cards, and each card bears one of two possible questions with two possible outcomes:

`Q1:`Do you have a nonsensitive attribute $A^c$? (with probability $\theta_2$).
`Q2:`Do you have a sensitive attribute $A$? (with probability $1-\theta_2$).

Assume that a simple random sample with the replacement (SRSWR) of $n$ respondents is selected from the population of interest. Each respondent is to report the first outcome of $R_1$ if he or she has a sensitive attribute $A$ and the second outcome of $R_2$ if he or she has no sensitive attribute $A$. The probability of a "yes" answer $\theta_K^*$ is given by

$$\theta_K^* = \pi\theta_1 + (1-\pi)\theta_2. \tag{4}$$

Let $n\hat{\theta}_K^*$ denote the number of "yes" responses in the sample of size $n$, and then the estimator $\hat{\pi}_K$ of $\pi$, the proportion of the population in the sensitive group, and its variance $V(\hat{\pi}_K)$ are given by

$$\hat{\pi}_K = \frac{\hat{\theta}_K^* - \theta_2}{\theta_1 - \theta_2}, \theta_1 \neq \theta_2, \tag{5}$$

$$V(\hat{\pi}_K) = \frac{\theta_K^*(1-\theta_K^*)}{n(\theta_1-\theta_2)^2}. \tag{6}$$

Many studies have suggested and extended various models based on Warner's model. Chaudhuri & Mukerjee (1988) and Ryu, Hong & Lee (1993) organize and emphasize various RR models. Kim & Warde (2004) present a stratified RR model by using an optimal allocation method, and Kim & Elam (2005) extend it to a two-stage stratified RR model. Recently Chaudhuri (2015) summrizes the history of RRM, Tarray & Singh (2015) suggest the Poisson RRM for a rare sentitive attribute. Also, Su, Sedory & Singh (2015) suggest a new RR model compelling answers "yes" or "no" to each respondent according to his or her selection situation in a randomization device modified from Kuk's randomization device. However, Su et al. (2015) model estimates sensitive attributes by using simple random sampling, and therefore it is difficult to apply it to populations composed of several strata.

This paper considers the conditions to estimate the proportion of sensitive attributes of a population composed by a number of strata and extends the adjusted Kuk model by applying stratified sampling. The paper estimates sensitive parameters in the case of a known stratum size by taking proportional and optimal allocation methods into account. It then extends it to the case of an unknown stratum size by estimating sensitive parameters by applying stratified double sampling to the Su, Sedory, and Singh model and checking the two allocation methods. Finally, the paper compares the efficiency of the proposed model to that of the Su, Sedory and Singh model and the stratified Kuk model in terms of the estimator variance.

## 2. An RR Model Using a Modified Kuk's Randomization Device

Su et al. (2015) estimate the proportion of sensitive attributes by suggesting an adjusted RR model that modified Kuk's. The modified Kuk model suggested in Su et al. (2015) applies the modified Kuk's randomization device to respondents selected by simple random sampling with replacement (SRSWR).

Each respondent in a sample of $n$ respondents is provided with two randomization devices $D_1$ and $D_2$. The randomization device $D_1$ consists of a deck of cards, and each card bears one of the following two statements: (1) use the randomization device $F_1$ and (2) use the randomization device $F_1^c$ with probabilities $\theta_1$ and $(1-\theta_1)$, respectively. Similarly, the randomization device $D_2$ consists of a deck of cards, and each card bears one of the following two statements: (1) use the randomization device $F_2$ and (2) use the randomization device $F_2^c$ with probabilities $\theta_2$ and $(1-\theta_2)$, respectively. Each respondent is instructed to use the first device $D_1$ if he or she has the sensitive attribute $A$ and the second device $D_2$ if he or she has the nonsensitive attribute $A^c$ . The device $F_1$ mentioned in the first outcome of the device $D_1$ consists of two possible mutually exclusive statements: (1) say "yes" and (2) say "no" with probabilities $P_1$ and $(1-P_1)$, respectively. The device $F_1^c$ mentioned in the second outcome of the device $D_1$ also consists of two possible mutually exclusive statements: (1) say "yes" and (2) say "no" with probabilities $T_1$ and $(1-T_1)$, respectively. Similarly, the device $F_2$ mentioned in the first outcome of the device $D_2$ consists of two possible mutually exclusive statements: (1) say "yes" and (2) say "no" with probabilities $P_2$ and $(1-P_2)$, respectively. The device $F_2^c$ mentioned in the second outcome of the device $D_2$ also consists of two possible mutually exclusive statements: (1) say "yes" and (2) say "no" but with probabilities $T_2$ and $(1-T_2)$, respectively.

In the adjusted Kuk RR model, the probability of a "yes" answer is given by

$$\begin{aligned}\theta_c^* &= \pi\left[\theta_1 P_1 + (1-\theta_1)T_1\right] + (1-\pi)\left[\theta_2 P_2 + (1-\theta_2)T_2\right]\\ &= \pi\left[\theta_1(P_1-T_1) - \theta_2(P_2-T_2) + (T_1-T_2)\right] + \theta_2 P_2 + (1-\theta_2)T_2,\end{aligned} \tag{7}$$

where $\pi$ is the population proportion of sensitive attributes.

The estimator $\hat{\pi}_c$ of the population proportion of sensitive attributes is

$$\hat{\pi}_c = \frac{\hat{\theta}_c^* - \theta_2 P_2 - (1-\theta_2)T_2}{\theta_1(P_1-T_1) - \theta_2(P_2-T_2) + (T_1-T_2)}, \tag{8}$$

where $\hat{\theta}_c^* = x/n$ is the observed proportion of "yes" answers.

The variance of the proposed estimator $\hat{\pi}_c$ is given as follows:

$$V(\hat{\pi}_c) = \frac{\theta_c^*(1-\theta_c^*)}{n\left[\theta_1(P_1-T_1) - \theta_2(P_2-T_2) + (T_1-T_2)\right]^2}. \tag{9}$$

## 3. A Stratified Kuk Randomization Device

This section considers the estimation of the proportion of sensitive attributes by using a stratified Kuk randomization device and checks the allocation method when the population consists of a number of strata and the size of each stratum is known. Let the population of size $N$ be divided into disjointed $L$ strata of size $N_h (h = 1, 2, \ldots, L)$ each in the stratum $h$. Then $n_h (n = \sum_{h=1}^{L} n_h)$ respondents are selected by the SRSWR and asked to answer "yes" or "no" according to the modified Kuk randomization device.

Each respondent in stratum $h$ is provided with two randomization devices $D_{h1}$ and $D_{h2}$. The randomization device $D_{h1}$ consists of a deck of cards, and each card bears one of the following two statements: (1) use the randomization device $F_{h1}$ and (2) use randomization device $F_{h1}^c$ with probabilities $\theta_{h1}$ and $(1-\theta_{h1})$, respectively. Similarly, the randomization device $D_{h2}$ consists of a deck of cards, and here each card bears one of the following two statements: (1) use the randomization device $F_{h2}$ and (2) use the randomization device $F_{h2}^c$ with probabilities $\theta_{h2}$ and $(1-\theta_{h1})$, respectively. Each respondent in stratum $h$ is instructed to use the first device $D_{h1}$ if he or she has the sensitive attribute $A_h$ and the second device $D_{h2}$ if he or she has no sensitive attribute $A_h$ . The device $F_{h1}$ mentioned in the first outcome $D_{h1}$ consists of two possible mutually exclusive statements: (1) say "yes" and (2) say "no" with probabilities $P_{h1}$ and $(1-P_{h1})$, respectively. The device $F_{h1}^c$ mentioned in the second outcome $D_{h1}$ consists of two possible mutually exclusive statements: (1) say "yes" and (2) say "no" with probabilities $T_{h1}$ and $(1-T_{h1})$, respectively. Similarly, the device $F_{h2}$ mentioned in the first outcome $D_{h2}$ consists of two possible mutually exclusive statements: (1) say "yes" and (2) say "no" with probabilities $P_{h2}$ and $(1-P_{h2})$, respectively. The device $F_{h2}^c$ mentioned in the second outcome $D_{h2}$ consists of two possible mutually exclusive statements: (1) say "yes" and (2) say "no" with probabilities $T_{h2}$ and $(1-T_{h2})$, respectively. A pictorial representation of this forced RR model is given in Figure 1.

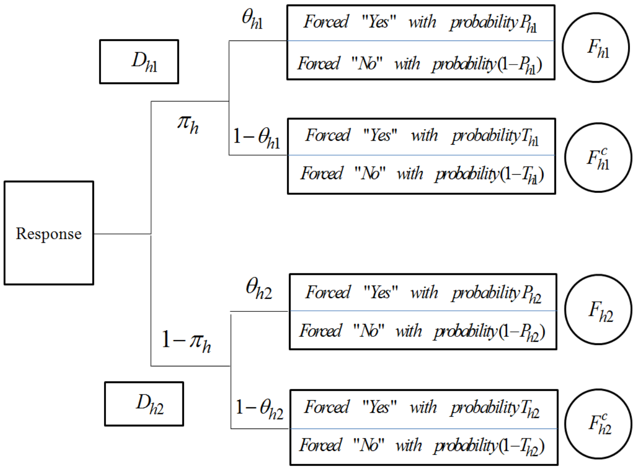
$\theta_{h1}$
$D_{h1}$
Forced "Yes" with probability $P_{h1}$
Forced "No" with probability $(1-P_{h1})$
$F_{h1}$
$\pi_h$
$1-\theta_{h1}$
Forced "Yes" with probability $T_{h1}$
Forced "No" with probability $(1-T_{h1})$
$F_{h1}^c$
Response
$\theta_{h2}$
Forced "Yes" with probability $P_{h2}$
Forced "No" with probability $(1-P_{h2})$
$F_{h2}$
$1-\pi_h$
$D_{h2}$
$1-\theta_{h2}$
Forced "Yes" with probability $T_{h2}$
Forced "No" with probability $(1-T_{h2})$
$F_{h2}^c$

FIGURE 1: A stratified Kuk forced randomized response model.

From these RR procedures, the probability of a "yes" answer is given by

$$\begin{aligned}\theta_{hc}^* &= \pi_h\left[\theta_{h1}P_{h1}+(1-\theta_{h1})T_{h1}\right]+(1-\pi_h)\left[\theta_{h2}P_{h2}+(1-\theta_{h2})T_{h2}\right]\\ &= \pi_h\left[(\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\right]+\theta_{h2}P_{h2},\\ &+(1-\theta_{h2})T_{h2}\end{aligned} \tag{10}$$

where $\pi_h$ is the population proportion of sensitive attributes in stratum $h$.

Let $X_h$ be the number of "yes" responses in the SRSWR sample of $n_h$ in the stratum $h$. Then $X_h$ follows a binomial distribution with parameters $n_h$ and $\theta^*_{hc}$, that is, $B(n_h, \theta^*_{hc})$. Therefore, the probability of observing $x_h$ "yes" answers out of $n_h$ responses is given by

$$P(X_h = x_h) = \binom{n_h}{x_h} (\theta^*_{hc})^{x_h} \, (1 - \theta^*_{hc})^{n_h - x_h}$$

The log-likelihood function is given by

$$\log P(X_h = x_h) = \log \binom{n_h}{x_h} + x_h \log{(\theta^*_{hc})} + (n_h - x_h) \log{(1 - \theta^*_{hc})}$$

Setting

$$\frac{\partial \log P(X_h = x_h)}{\partial \theta^*_{hc}} = 0$$

gives the maximum likelihood estimator (MLE) $\hat{\theta}^*_{hc}$ of $\theta^*_{hc}$ as follows:

$$\hat{\theta}^*_{hc} = \frac{x_h}{n_h} \tag{11}$$

Therefore, the estimator $\hat{\pi}_{hc}$ of the population proportion of sensitive attributes in stratum $h$ is

$$\hat{\pi}_{hc} = \frac{\hat{\theta}^*_{hc} - \theta_{h2} P_{h2} - (1 - \theta_{h2}) T_{h2}}{\theta_{h1}(P_{h1} - T_{h1}) - \theta_{h2}(P_{h2} - T_{h2}) + (T_{h1} - T_{h2})}, \tag{12}$$

where $\hat{\theta}^*_{hc} = x_h/n_h$ is the observed proportion of "yes" answers in the sample of $h$.

**Theorem 1.** *The stratified estimator $\hat{\pi}_{st}$ is an unbiased estimator of $\pi$:*

$$\begin{aligned} \hat{\pi}_{st} &= \sum_{h=1}^{L} W_h \hat{\pi}_{hc} \\ &= \sum_{h=1}^{L} W_h \frac{\hat{\theta}^*_{hc} - \theta_{h2} P_{h2} - (1 - \theta_{h2}) T_{h2}}{\theta_{h1}(P_{h1} - T_{h1}) - \theta_{h2}(P_{h2} - T_{h2}) + (T_{h1} - T_{h2})} \end{aligned} \tag{13}$$

*where $W_h = \frac{N_h}{N}$.*

***Proof*.** Because $E(\hat{\theta}^*_{hc}) = \theta^*_{hc}$, it is easy to show that

$$\begin{aligned}
E(\hat{\pi}_{st}) &= E\left[\sum_{h=1}^{L} W_h \frac{\hat{\theta}^*_{hc} - \theta_{h2}P_{h2} - (1-\theta_{h2})T_{h2}}{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})}\right] \\
&= \sum_{h=1}^{L} W_h \frac{E(\hat{\theta}^*_{hc}) - \theta_{h2}P_{h2} - (1-\theta_{h2})T_{h2}}{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})} \\
&= \sum_{h=1}^{L} W_h \frac{\theta^*_{hc} - \theta_{h2}P_{h2} - (1-\theta_{h2})T_{h2}}{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})} \\
&= \sum_{h=1}^{L} W_h \frac{\pi_h[\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})]}{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})} \\
&+ \sum_{h=1}^{L} W_h \frac{\theta_{h2}P_{h2} + (1-\theta_{h2})T_{h2} - \theta_{h2}P_{h2} - (1-\theta_{h2})T_{h2}}{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})} \\
&= \sum_{h=1}^{L} W_h \pi_h \\
&= \pi.
\end{aligned}$$

□

**Theorem 2.** *The variance of the proposed estimator* $\hat{\pi}_{st}$ *is given as follows:*

$$V(\hat{\pi}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{\theta^*_{hc}(1-\theta^*_{hc})}{n_h\left[\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})\right]^2}. \tag{14}$$

***Proof*.** Because $X_h \sim B(n_h, \theta^*_{hc})$ and is independent, it is easy to show that

$$\begin{aligned}
V(\hat{\pi}_{st}) &= V\left[\sum_{h=1}^{L} W_h \frac{\hat{\theta}^*_{hc} - \theta_{h2}P_{h2} - (1-\theta_{h2})T_{h2}}{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})}\right] \\
&= \sum_{h=1}^{L} W_h^2 \frac{V(\hat{\theta}^*_{hc})}{\left[\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})\right]^2} \\
&= \sum_{h=1}^{L} W_h^2 \frac{\theta^*_{hc}(1-\theta^*_{hc})}{n_h\left[\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})\right]^2}
\end{aligned}$$

□

**Theorem 3.** *An unbiased estimator of the variance of the proposed estimator* $\hat{\pi}_{st}$ *is*

$$\hat{V}(\hat{\pi}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{\hat{\theta}^*_{hc}(1-\hat{\theta}^*_{hc})}{(n_h-1)\left[\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})\right]^2}. \tag{15}$$

***Proof*.** This is obvious because $E[\hat{\theta}^*_{hc}(1-\hat{\theta}^*_{hc})] = (n_h - 1)V(\hat{\theta}^*_{hc})$:

$$\begin{aligned}
E\left[\hat{V}(\hat{\pi}_{st})\right] &= E\left[\sum_{h=1}^{L} W_h^2 \frac{\hat{\theta}^*_{hc}(1-\hat{\theta}^*_{hc})}{(n_h-1)\left[\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\right]^2}\right] \\
&= \sum_{h=1}^{L} W_h^2 \frac{E\left[\hat{\theta}^*_{hc}(1-\hat{\theta}^*_{hc})\right]}{(n_h-1)\left[\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\right]^2} \\
&= \sum_{h=1}^{L} W_h^2 \frac{(n_h-1)V(\hat{\theta}^*_{hc})}{(n_h-1)\left[\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\right]^2} \\
&= \sum_{h=1}^{L} W_h^2 \frac{\theta^*_{hc}(1-\theta^*_{hc})}{n_h\left[\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\right]^2} \\
&= V(\hat{\pi}_{st}).
\end{aligned}$$

□

Now, consider proportional and optimal allocation methods to allocate the overall sample of $n$ to each stratum of $n_h$ and check the variance in each case. In stratified sampling, values of sample sizes $n_h$ in respective strata are chosen by the sampler. If the stratum size $N_h$ is known but the variance of each stratum is not known, then the proportional allocation method is useful. In proportional allocation, $n_h = n(N_h/N)$, and the variance of $\hat{\pi}_{st}$ is given by

$$V(\hat{\pi}_{st}) = \frac{1}{n}\sum_{h=1}^{L} W_h \frac{\theta^*_{hc}(1-\theta^*_{hc})}{\left[\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\right]^2}. \tag{16}$$

The optimal allocation method determines $n_h$ to minimize $V(\hat{\pi}_{st})$ for a specified cost or the cost of a specified value of $V(\hat{\pi}_{st})$. Let the cost function be

$$C = c_0 + \sum_{h=1}^{L} c_h n_h \tag{17}$$

where $c_0$ is the overhead cost and $c_h$ is the cost per unit.

In the optimal allocation method, the stratum sample size $n_h$ and the minimum variance of $\hat{\pi}_{st}$ are respectively given as

$$n_h = n\frac{W_h\sqrt{\dfrac{\theta^*_{hc}(1-\theta^*_{hc})}{[\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})]^2}}\Big/\sqrt{c_h}}{\sum_{h=1}^{L} W_h\sqrt{\dfrac{\theta^*_{hc}(1-\theta^*_{hc})}{[\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})]^2}}\Big/\sqrt{c_h}} \tag{18}$$

$$\begin{aligned}
V(\hat{\pi}_{st(o)}) = \frac{1}{n}\sum_{h=1}^{L} W_h\sqrt{c_h}&\sqrt{\frac{\theta^*_{hc}(1-\theta^*_{hc})}{[\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})]^2}} \\
\times\sum_{h=1}^{L}\frac{W_h}{\sqrt{c_h}}&\sqrt{\frac{\theta^*_{hc}(1-\theta^*_{hc})}{[\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})]^2}}.
\end{aligned} \tag{19}$$

## 4. Stratified Double Estimation of Sensitive Attributes Using a Stratified Kuk Randomization Device

This section considers the estimation of the proportion of sensitive attributes by using a stratified Kuk randomization device and checks the allocation method when the population consists of a number of strata but there is no information on the size of each stratum. If there is a lack of information on stratum size in stratified sampling, then stratified double sampling is useful. If there is a lack of information on stratum size, then it can be obtained from the first sample, and the estimation of sensitive attributes can be made by using the stratified Kuk randomization device. If a population of size $N$ consists of $L$ strata, then the first sample of $n'$ respondents is selected by the SRSWR, and they are asked to directy answer the question “Are you in stratum?”. Then the first sample is classified into $h$ strata of size $n'_h$, and two proportions $W_h$ and $w_h$ are defined as follows:

$$W_h = \frac{N_h}{N} \text{ : The proportion of the population falling into stratum } h,$$

$$w_h = \frac{n'_h}{n'} \text{ : The proportion of the first sample falling into stratum } h.$$

where $w_h$ is an unbiased estimator of $W_h$ for $h = 1, 2, \ldots, L$.

The second sample is a stratified sample in stratum $h$. Here $n_h(n = \sum_{h=1}^{L} n_h)$ respondents are selected by the SRSWR from the first sample $n'_h$; they are then asked to answer “yes” or “no” according to the stratified Kuk randomization device, as in Section 2. The stratified estimator $\hat{\pi}_{std}$ of the population proportion of sensitive attributes can be obtained from these procedures as follows:

$$\hat{\pi}_{std} = \sum_{h=1}^{L} w_h \frac{\hat{\theta}^*_{hc} - \theta_{h2}P_{h2} - (1-\theta_{h2})T_{h2}}{\theta_{h1}(P_{h1} - T_{h1}) - \theta_{h2}(P_{h2} - T_{h2}) + (T_{h1} - T_{h2})}, w_h = \frac{n'_h}{n'}. \tag{20}$$

**Theorem 4.** *The stratified estimator $\hat{\pi}_{std}$ is an unbiased estimator of $\pi$.*

***Proof*.** Because $E(\hat{\theta}^*_{hc}) = \theta^*_{hc}$, it is easy to show that

$$\begin{aligned}
E\left[\hat{\pi}_{std}\right] &= E_1\left[E_2\left(\sum_{h=1}^{L} w_h \frac{\hat{\theta}^*_{hc} - \theta_{h2}P_{h2} - (1-\theta_{h2})T_{h2}}{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})}\right)|w_h\right] \\
&= E_1\left[\sum_{h=1}^{L} w_h \frac{E_2(\hat{\theta}^*_{hc}) - \theta_{h2}P_{h2} - (1-\theta_{h2})T_{h2}}{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})}\right] \\
&= E_1\left[\sum_{h=1}^{L} w_h \frac{\theta^*_{hc} - \theta_{h2}P_{h2} - (1-\theta_{h2})T_{h2}}{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})}\right] \\
&= E_1\left[\sum_{h=1}^{L} w_h \frac{\pi_h[\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})]}{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})}\right. \\
&\left. + \sum_{h=1}^{L} w_h \frac{\theta_{h2}P_{h2} + (1-\theta_{h2})T_{h2} - \theta_{h2}P_{h2} - (1-\theta_{h2})T_{h2}}{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})}\right] \\
&= E_1\left[\sum_{h=1}^{L} w_h \pi_h\right] \\
&= \sum_{h=1}^{L} W_h \pi_h = \pi
\end{aligned}$$

□

**Theorem 5.** *The variance of the proposed estimator $\hat{\pi}_{std}$ is given as follows:*

$$\begin{aligned}
V(\hat{\pi}_{std}) &= \frac{1}{n'}\left[\sum_{h=1}^{L} W_h \left(\frac{\theta^*_{hc}(1-\theta^*_{hc})}{[\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})]^2}\right)\right. \\
&\left. + \sum_{h=1}^{L} W_h(\pi_h - \pi)^2\right] \\
&+ \sum_{h=1}^{L} \frac{W_h}{n'}\left(\frac{1}{v_h} - 1\right)\left(\frac{\theta^*_{hc}(1-\theta^*_{hc})}{[\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})]^2}\right)
\end{aligned} \tag{21}$$

*where $0 \leq v_h = n_h/n'_h \leq 1$ is a fixed constant.*

***Proof*.** If $\hat{\pi}'_h$ is written as an estimator of sensitive attributes obtained from the first sample $n'_h$ of the stratum $h$ and $\hat{\pi}_{std}$ is redefined as a function of $\hat{\pi}'_h$, then $\hat{\pi}_{std}$ can be expressed as follows:

$$\hat{\pi}_{std} = \sum_{h=1}^{L} w_h \hat{\pi}_h = \sum_{h=1}^{L} w_h \hat{\pi}'_h + \sum_{h=1}^{L} w_h(\hat{\pi}_h - \hat{\pi}'_h).$$

The variance of the first term on the right-hand side is

$$
\begin{aligned}
&V_1 E_2\left(\sum_{h=1}^{L} w_h \hat{\pi}'_h\right)\\
&= \frac{1}{n'}\left[\sum_{h=1}^{L} W_h\left(\frac{\theta^*_{hc}(1-\theta^*_{hc})}{\{\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\}^2}\right)\right.\\
&\left.+\sum_{h=1}^{L} W_h(\pi_h-\pi)^2\right],
\end{aligned}
$$

and the second term on the right-hand side is

$$
\begin{aligned}
&E_1\left[V_2\left(\sum_{h=1}^{L} w_h(\hat{\pi}_h-\hat{\pi}'_h)\right)\right]\\
&= E_1\left[\sum_{h=1}^{L}\left(\frac{1}{n_h}-\frac{1}{n'_h}\right)w_h^2\left(\frac{\theta^*_{hc}(1-\theta^*_{hc})}{\{\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\}^2}\right)\right]\\
&= E_1\left[\sum_{h=1}^{L}\frac{w_h}{n'}\left(\frac{1}{v_h}-1\right)\left(\frac{\theta^*_{hc}(1-\theta^*_{hc})}{\{\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\}^2}\right)\right]\\
&= \sum_{h=1}^{L}\frac{W_h}{n'}\left(\frac{1}{v_h}-1\right)\left(\frac{\theta^*_{hc}(1-\theta^*_{hc})}{\{\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\}^2}\right),
\end{aligned}
$$

because $n_h = v_h n'_h = v_h w_h n'$.

Here (21) is obtained from these equations. □

Now consider proportional and optimal allocation methods to allocate the overall sample of $n$ to each stratum of $n'_h$ and check the variance in each case. If the stratum size $N_h$ is known but the variance of each stratum is not known, then the proportional allocation method is useful. In the proportional allocation method, if $n'$ and $n'_h$ are used instead of $N$ and $N_h$, then $n_h = n(n'_h/n')$ and the variance of $\hat{\pi}_{std(p)}$ is given by

$$
\begin{aligned}
V(\hat{\pi}_{std(p)}) &= \frac{1}{n'}\sum_{h=1}^{L} W_h(\hat{\pi}_h-\pi)^2\\
&+\frac{1}{n}\sum_{h=1}^{L} W_h\left(\frac{\theta^*_{hc}(1-\theta^*_{hc})}{\{\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\}^2}\right).
\end{aligned}
\tag{22}
$$

The optimal allocation method determines $n'$ and $v_h$ to minimize $V(\hat{\pi}_{std})$ for a specified cost. Let the cost function be

$$C = c'n' + \sum_{h=1}^{L} c_h n_h. \tag{23}$$

where $c'$ is the total cost for the first sample and $c_h$ is the cost per unit.

Then the expected value of (23) must be minimized to obtain optimum values of $n'$ and $v_h$ because $n_h$ is a random variable. The expected value of $C$ is given by

$$E(C) = C^* = c'n' + \sum_{h=1}^{L} c_h E(n_h) = c'n' + n' \sum_{h=1}^{L} c_h v_h W_h. \tag{24}$$

Use the Cauchy-Schwarz inequality to obtain $v_h$ that minimizes the product $V(\hat{\pi}_{std})E(C)$:

$$v_h = \sqrt{\frac{c'}{c_h} \frac{\frac{\theta^*_{hc}(1-\theta^*_{hc})}{\{\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\}^2}}{\sum_{h=1}^{L} W_h(\pi_h - \pi)^2}} \tag{25}$$

Substituting $v_h$ into (24) gives the optimum value of $n'$ as follows:

$$n' = \frac{C^*}{c' + \sum_{h=1}^{L} c_h W_h \sqrt{\frac{c'}{c_h} \frac{\frac{\theta^*_{hc}(1-\theta^*_{hc})}{\{\theta_{h1}(P_{h1}-T_{h1})-\theta_{h2}(P_{h2}-T_{h2})+(T_{h1}-T_{h2})\}^2}}{\sum_{h=1}^{L} W_h(\pi_h-\pi)^2}}} \tag{26}$$

Therefore, the minimum variance of $\hat{\pi}_{std(o)}$ is

$$\begin{aligned} V(\hat{\pi}_{std(o)}) = \frac{1}{C^*} \Bigg[ & \sqrt{c'} \sqrt{\sum_{h=1}^{L} W_h(\pi_h - \pi)^2} \\ & + \sum_{h=1}^{L} W_h \sqrt{\frac{\theta^*_{hc}(1-\theta^*_{hc})}{\{\theta_{h1}(P_{h1}-T_{h1}) - \theta_{h2}(P_{h2}-T_{h2}) + (T_{h1}-T_{h2})\}^2}} \sqrt{c_h} \Bigg]^2 . \end{aligned} \tag{27}$$

# 5. Efficiency Comparison

## 5.1. Stratified Estimation vs. Su, Sedory and Singh Estimation

In the Su et al. (2015) model, the variance of the estimator $\hat{\pi}_c$ of the sensitive attribute $\pi$ is

$$V(\hat{\pi}_c) = \frac{\theta^*_c(1-\theta^*_c)}{n\left[\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)\right]^2}, \tag{28}$$

where $\theta^*_c = \pi[\theta_1(P_1 - T_1) - \theta_2(P_2 - T_2) + (T_1 - T_2)] + \theta_2 P_2 + (1-\theta_2)T_2$.

Here the relative efficiency used ($RE$) to compare the efficiency of two models:

$$RE = \frac{V(\hat{\pi}_c)}{V(\hat{\pi}_{st})}.$$

Values of $RE$ greater than 1 indicate that the estimator obtained using the proposed stratified estimation method is more efficient than the estimator in Su et al. (2015). To calculate $RE$ empirically, it is assumed that the population has two strata and $W_1 = 0.6, W_2 = 0.4$ and $W_1 = 0.7, W_2 = 0.3$ for $N = \sum_{h=1}^{2} N_h = 10,000$. It is also assumed that $\theta_1 = \theta_{11} = \theta_{21} = 0.7, \theta_2 = \theta_{12} = \theta_{22} = 0.2$ and $0.1 \leq \pi = \pi_1 = \pi_2 \leq 0.3$. Tables 1 and 2 show the frequency of $RE > 1$ when values of $P_{h1}, P_{h2}, \ T_{h1}, T_{h2}, h = 1, 2$ increase from 0.7 to 0.9 by 0.1 for $\pi = \pi_1 = \pi_2 = 0.1$. The total number of iterations is $14, 348, 907$, and the cases of $RE > 1$ are $1, 711, 724$ and $1, 734, 572$ for $\pi = \pi_1 = \pi_2 = 0.1$, $W_1 = 0.6, W_2 = 0.4$ and $W_1 = 0.7, W_2 = 0.3$ (refer to Tables 1 and 2).

Table 1: Frequency of $RE > 1$ for $\pi = \pi_1 = \pi_2 = 0.1$ and $W_1 = 0.6, W_2 = 0.4$ (%).

| | $T_1 = T_{11} = T_{21}$ | | | | $T_2 = T_{12} = T_{22}$ | | | |
|---|---|---|---|---|---|---|---|---|
| $P_1 = P_2$ | 0.7 | 0.8 | 0.9 | Total | 0.7 | 0.8 | 0.9 | Total |
| 0.7 | 114,147 | 214,249 | 220,736 | 549,132 | 116,726 | 263,613 | 202,616 | 582,955 |
| | (20.79) | (39.02) | (40.2) | (100.00) | (20.02) | (45.22) | (34.76) | (100.00) |
| 0.8 | 229,525 | 182,722 | 223,052 | 635,299 | 207,221 | 201,021 | 165,003 | 573,245 |
| | (36.13) | (28.76) | (35.11) | (100.00) | (36.15) | (35.07) | (28.78) | (100.00) |
| 0.9 | 216,002 | 207,069 | 104,222 | 527,293 | 236,356 | 253,090 | 66,078 | 555,524 |
| | (40.96) | (39.27) | (19.77) | (100.00) | (42.55) | (45.56) | (11.89) | (100.00) |
| Total | 559,674 | 604,040 | 548,010 | 1,711,724 | 560,303 | 717,724 | 433,697 | 1,711,724 |

Table 2: Frequency of $RE > 1$ for $\pi = \pi_1 = \pi_2 = 0.1$ and $W_1 = 0.7, W_2 = 0.3$ (%).

| | $T_1 = T_{11} = T_{21}$ | | | | $T_2 = T_{12} = T_{22}$ | | | |
|---|---|---|---|---|---|---|---|---|
| $P_1 = P_2$ | 0.7 | 0.8 | 0.9 | Total | 0.7 | 0.8 | 0.9 | Total |
| 0.7 | 116,692 | 215,194 | 225,179 | 557,065 | 118,509 | 267,839 | 202,586 | 588,934 |
| | (20.95) | (38.63) | (40.42) | (100.00) | (20.12) | (45.48) | (34.4) | (100.00) |
| 0.8 | 232,422 | 186,760 | 226,949 | 646,131 | 210,034 | 205,524 | 165,167 | 580,725 |
| | (35.97) | (28.90) | (35.12) | (100.00) | (36.17) | (35.39) | (28.44) | (100.00) |
| 0.9 | 218,158 | 208,278 | 104,940 | 531,376 | 239,597 | 256,742 | 68,574 | 564,913 |
| | (41.06) | (39.2) | (19.75) | (100.01) | (42.41) | (45.45) | (12.14) | (100.00) |
| Total | 567,272 | 610,232 | 557,068 | 1,734,572 | 568,140 | 730,105 | 436,327 | 1,734,572 |

Based on Table 1, if $P_1 = P_2 = 0.7$ and $T_2 = T_{12} = T_{22} = 0.7$ , then the percentage of $RE > 1$ is 20.79%, and if $T_2 = T_{12} = T_{22} = 0.7$, then the percentage of $RE > 1$ is 20.02%. In the case of some fixed $P_1 = P_2 = 0.7$, if $T_1, T_2$ increase from 0.7 to 0.9, then the percentage of $RE > 1$ increases. However, if $P_1, P_2$ or $T_1, T_2$ increase from 0.7 to 0.9, then the percentage of $RE > 1$ decreases. In addition, if selection probabilities $P_1, P_2$ and $T_1, T_2$ have the same value, then the percentage of $RE > 1$, that is, that of diagonal cells, will have the lowest value of any other off-diagonal cells.

Figures 2 and 3 show that section probabilities $P_1 = P_2 = 0.7$ to 0.9 and $T_2 = T_{12} = T_{22} = 0.7$ to 0.9 for each stratum reduce the percentage of $RE > 1$. If $\pi = \pi_1 = \pi_2 = 0.2$ or 0.3, the percentage of $RE > 1$ has the same pattern as that in Figures 2 and 3. Figure 4 shows that $RE > 1$ increases for $\pi = \pi_1 = \pi_2 = 0.1$ to 0.3 and decreases for $\pi = \pi_1 = \pi_2 = 0.7$ to 0.9 for strata sizes $W_1 = 0.7, W_2 = 0.3$ and $W_1 = 0.6, W_2 = 0.4$, respectively. In addition, the percentage of $RE > 1$ if $W_1 = 0.7, W_2 = 0.3$ are greater than $W_1 = 0.6, W_2 = 0.4$. That is, if the size of the stratum varies across strata, then the relative efficiency of the proposed estimator is more efficient than Su et al.'s (2015) estimator.

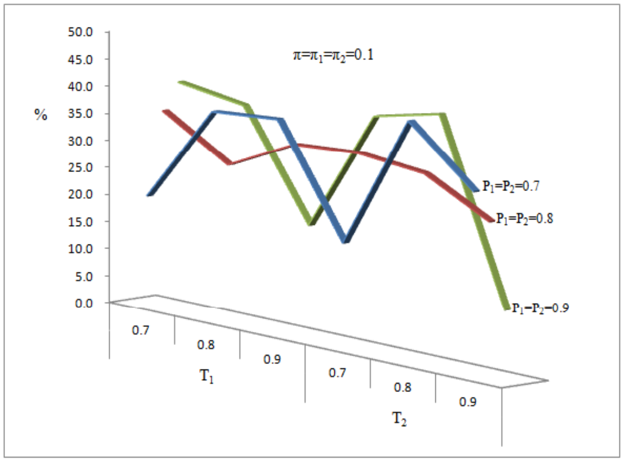

FIGURE 2: Percentages of $RE > 1$ for $\pi = \pi_1 = \pi_2 = 0.1$ and $W_1 = 0.6, W_2 = 0.4$.

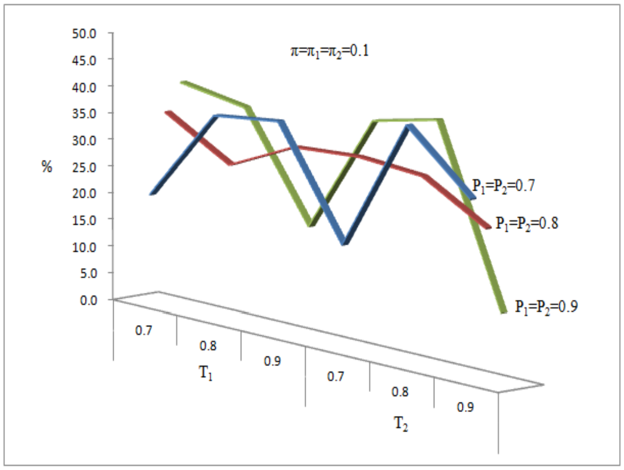

FIGURE 3: Percentages of $RE > 1$ for $\pi = \pi_1 = \pi_2 = 0.1$ and $W_1 = 0.7, W_2 = 0.3$.

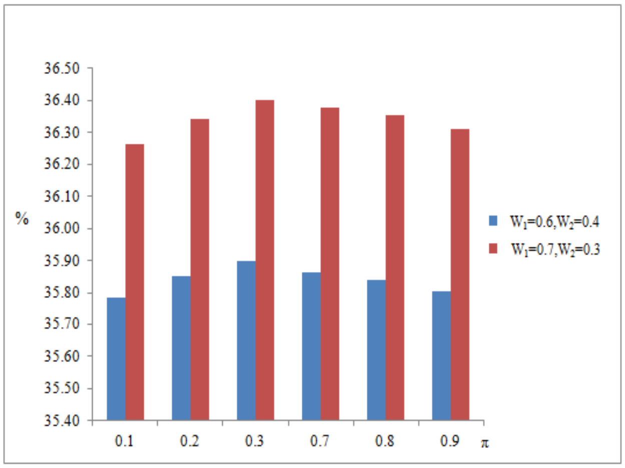

Figure 4: Percentages of $RE > 1$ for $\pi = \pi_1 = \pi_2 = 0.1$ to $0.9$, $W_1 = 0.7, W_2 = 0.3$ and $W_1 = 0.6, W_2 = 0.4$.

### 5.2. Stratified Estimation vs. Stratified Double Estimation

The difference between (16), the variance of the stratified estimator, and (22): the variance of the stratified double estimator, is

$$\frac{1}{n'} \sum_{h=1}^{L} W_h (\hat{\pi}_h - \pi)^2. \tag{29}$$

The increment of the variance in stratified double sampling is due to an unknown stratum size obtained in the process of forming an estimator.

## 6. Conclusions

This paper estimate sensitive attributes of a population composed of a number of strata by applying stratified sampling to the Su, Sedory and Singh model. The paper estimates sensitive parameters in the case where stratum size is known by taking proportional and optimal allocation methods into account; this is then to the case of an unknown stratum size, for which sensitive parameters are estimated by applying stratified double sampling to the modified Kuk model and the two allocation methods are checked. The paper compares the efficiency of the proposed stratified Kuk model to that of the Su, Sedory and Singh model in terms of estimator variance. The results of the numerical study indicate that the proposed estimator is more efficient than the Su, Sedory and Singh model for different sizes of strata. In the proposed model, $RE > 1$ is guaranteed to be more than 35% in all cases with various parameters.

# Acknowledgements

This research was supported by the Basic Science Research Program of the National Research Foundation of Korea (NRF), which is funded by the Ministry of Science, ICT, and Future Planning (2015R1A2A2A 01003699). The corresponding author of this research is Assistant professor Son Chang-Kyoon.

[Received: October 2015 — Accepted: August 2016]

# References

Chaudhuri, A. (2015), 'Fifty years gone by', *Model Assisted Statistics and Applications* **10**(4), 277–282.

Chaudhuri, A. & Mukerjee, R. (1988), *Randomized Response: Theory and Techniques*, Marcel Dekker, Inc., New York.

Kim, J. M. & Elam, M. E. (2005), 'A Two-Stage stratified Warner's randomized Response Model Using Optimal Allocation', *Metrika* **61**, 1–7.

Kim, J. M. & Warde, W. D. (2004), 'A stratified Warner's randomized response model', *Journal of Statistical Planning and Inference* **120**, 155–165.

Kuk, A. Y. C. (1990), 'Asking sensitive questions indirectly', *Biometrika* **77**, 436–438.

Ryu, J. B., Hong, K. H. & Lee, G. S. (1993), *Randomized Response Model*, Freedom Academy, Seoul.

Su, S. C., Sedory, S. A. & Singh, S. (2015), 'Kuk's model adjusted for protection and efficiency', *Sociological Methods & Research* **43**(3), 534–551.

Tarray, T. & Singh, H. (2015), 'A randomized response model for estimating a rare sensitive attribute in stratified sampling using Poisson distribution', *Model assisted Statistics & Applications* **10**(5), 361–384.

Warner, S. L. (1965), 'Randomized Response; A Survey Technique for Eliminating Evasive Answer Bias', *Journal of the American Statistical Association* **60**, 63–69.